EXHIBIT C

AUTOMATIC SPEECH AND SPEAKER RECOGNITION

Advanced Topics

edited by

Chin-Hui Lee
Frank K. Soong
AT&T Bell Laboratories

Kuldip K. Paliwal
School of Microelectronic Engineering
Griffith University

KLUWER ACADEMIC PUBLISHERS
Boston / Dordrecht / London

Distributors for North, Central and South America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA
Telephone (781) 871-6600
Fax (781) 871-6528
E-Mail <kluwer@wkap.com>

Distributors for all other countries:
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS
Telephone 31 78 6392 392
Fax 31 78 6546 474
E-Mail services@wkap.nl>

 Electronic Services <http://www.wkap.nl>

Library of Congress Cataloging-in-Publication

Automatic speech and speaker recognition : advanced topics / edited by
Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal.
p. cm. -- (The Kluwer international series in engineering and computer science : SECS 355)
Includes bibliographical references and index.
ISBN 0-7923-9706-1
1. Automatic speech recognition. I. Lee, Chin-Hui. II. Soong, Frank K. III. Paliwal, K.K. (Kuldip K.) IV. Series
TK7895.S65A98 1996 96-1588
006.4'54--dc 20 CIP

Printed on acid-free paper.

Printed in the United States of America

14

SPECTRAL DYNAMICS FOR SPEECH RECOGNITION UNDER ADVERSE CONDITIONS

Brian A. Hanson, Ted H. Applebaum and Jean-Claude Junqua

Speech Technology Laboratory, Panasonic Technologies, Inc.
Santa Barbara, California, USA

ABSTRACT

Significant improvements in automatic speech recognition performance have been obtained through front-end feature representations which exploit the time varying properties of speech spectra. Various techniques have been developed to incorporate "spectral dynamics" into the speech representation, including *temporal derivative features*, *spectral mean normalization* and, more generally, *spectral parameter filtering*. This chapter describes the implementation and interrelationships of these techniques and illustrates their use in automatic speech recognition under different types of adverse conditions.

1 INTRODUCTION

As speech recognition systems are applied in the real world, the kinds of variability encountered in the input speech signal become less predictable. Assumptions about the speech signal and the processes which distort it can be used to overcome specific types of adverse conditions. Recognition performance may suffer when these assumptions are violated. Hence, to obtain robust speech recognition, techniques which depend on assumptions of specific conditions are best avoided. This is the case, for example, for the use of the spectral dynamics of speech in recognition systems. Simple but general techniques which emphasize spectral changes have resulted in significant improvements to recognition performance, while reducing the sensitivity of the system to irrelevant sources of variability.

Accurate speech recognition requires that the signal representation preserve and enhance phonetically discriminating information in the speech signal. As noted by Picone [1], early recognition systems avoided, as unreliable, signal representations which emphasized the dynamics of the spectrum, despite the fact that rapid spectral changes are a major cue in phonetic discrimination. Improved speech modeling techniques, such as hidden Markov modeling, which can characterize the time-varying aspect of the signal, have enabled the routine use of dynamic features, which is nearly universal in current speech recognition systems.

This chapter begins by briefly describing the kinds of adverse conditions which affect speech signals and reviewing motivations for application of spectral dynamic features to compensate for such effects. The bulk of the chapter is given to a description of techniques for robust speech recognition based on spectral dynamics. These techniques, including temporal derivative features and spectral mean normalization, are interpreted from the point of view of filtering the sequence of spectral parameters in a particular domain. Comparative examples of recognition performance are given for clean and degraded (noisy and/or spectrally distorted) speech. The chapter concludes with a discussion of the properties of spectral dynamics which contribute to improved speech recognition performance.

2 ADVERSE CONDITIONS WHICH AFFECT SPEECH SIGNALS

Speech is a dynamic acoustic signal with many sources of variation. For example, the production of different phonemes involves a variety of movements of the speech articulators and there is much freedom in the timing and degree of vocal tract movements. Consequently, depending on a number of conditions, a speaker can greatly modify his speech production but still transmit the same linguistic message. The adverse conditions which can affect automatic speech recognition fall into three broad categories:

Talker and task variations can be induced by factors such as social context, speaking style, speaking rate, voice quality, stress due to the environmental conditions, cognitive load, and emotion. Stress, and more specifically the Lombard reflex (i.e. how a talker's speech changes in response to noise), has been the object of a number of studies (e.g. [2–4]). Spectral slope, energy distribution, formant frequencies, and cepstral norm have been found to be affected by the Lombard reflex. Moreover, acoustic differences between speech produced in noise-free conditions and speech produced in noise modify the intelligibility of speech (e.g. [4–6]).

The *acoustic environment* influences speech production but also may distort the speech signal. While at low signal-to-noise ratio (SNR) humans understand noisy speech reasonably well, performance of automatic speech recognition systems is reduced considerably. In a noisy environment, spectral valleys are more affected by noise than spectral peaks [7]. Furthermore, additive white-Gaussian noise reduces the norm of linear prediction cepstral vectors [8]. Non-stationary noise and other variabilities due to the environment are still not well handled by speech recognition systems.

After speech has been produced and before it reaches the listener, it may be distorted by *microphones and transmission channels.* The induced distortions act generally on the speech signal as convolutional noises. It was reported that the word accuracy of the SPHINX speech recognition system dropped from 85% to below 20% when a desk-top microphone was substituted for the close-talking microphone used in training [9]. Recently, many studies have focused on microphone-independence and robust techniques against channel distortions (e.g. [10–16]).

3 SPEECH RECOGNITION BASED ON SPECTRAL DYNAMICS

Various approaches have been tried for improving speech recognition in adverse conditions. Improvements have been obtained by techniques applied during feature estimation, matching, and language modeling (see reviews in [16–19]). Although each stage of the recognition process can contribute to robustness, it is essential that the front-end speech representation, which is the basis for all subsequent processing, be relatively insensitive to irrelevant sources of variation in the signal. Numerous techniques have attempted to realize such insensitivity by utilizing some form of temporal dynamics of the speech spectra. Before describing these techniques and the improvements in recognition obtained from them, we review some of the motivations for using spectral dynamics.

3.1 Motivations for Using Spectral Dynamic Features

Spectral transitions play an important role in human auditory perception. A wide variety of approaches have been investigated that enhance the robustness of the front-end analysis of automatic speech recognition by exploiting characteristics of the time variations, i.e. *dynamics*, of speech spectra [20–31].

Numerous experiments have demonstrated the importance of spectral dynamics in speech perception. The experiments reported by Furui [32] are particularly relevant since he compares results from a series of perceptual experiments to a spectral transition measure based on a spectral time derivative. These experiments focused on determining the time location of the "perceptual critical point" of a syllable, defined as the amount of truncation of a syllable where its identification rate dropped below 80%. Furui found that the perceptual critical point occurs in the same approximately 10 msec speech segment as the maximum spectral transition (as measured by a time-derivative-based transition measure). Furthermore, the spectral transitions were found to be important in both consonant *and* vowel perception, indicating that spectral dynamics are crucial for phone perception.

One form of spectral dynamics that can be directly examined is the first derivative of the "static" spectral feature. Static spectral feature, $S(f, n\Delta T)$,[1] as used here can refer to the results of any of the standard front-end analyses used in speech recognition, e.g. filterbank, Linear Prediction (LP) [33], Perceptually-based Linear Prediction (PLP) [34, 35], or mel-frequency cepstrum [36]. The first derivative can be approximated by taking the *first difference*, $D(f, n\Delta T)$, of the static spectral features from speech frames separated by a short time interval $2\delta_D$:

$$D(f, n\Delta T) = S(n\Delta T + \delta_D) - S(n\Delta T - \delta_D) \quad (1)$$

The characteristics of the temporal derivative approximation of eqn. 1 may be visually examined in a frequency versus time spectrogram-like display, as shown in Fig. 1. The top of Fig. 1a shows a standard wideband spectrogram from the utterance "nine". Immediately below this is a "spectrogram" based on index-weighted cepstral [37] coefficients derived from PLP analysis. This analysis was used for its spectral-peak emphasizing properties [38].

The bottom two spectrograms of Fig. 1a represent the positive and negative parts of the first difference feature, $D(f, n\Delta T)$. Unlike typical "static feature" spectrograms, temporal derivatives can have positive and negative components depending on the direction of the spectral changes over time, so on a gray-scale display it is necessary to plot them separately.

In the difference spectrograms, formant onsets are clearly highlighted by the "positive component", and formant offsets are emphasized in the "negative component". Much of the noise and other short-term artifacts present in the

[1] For simplicity, the frequency dependence of S will be omitted here and the time dependence shown as a function of frame number and frame step size, n and ΔT, respectively.

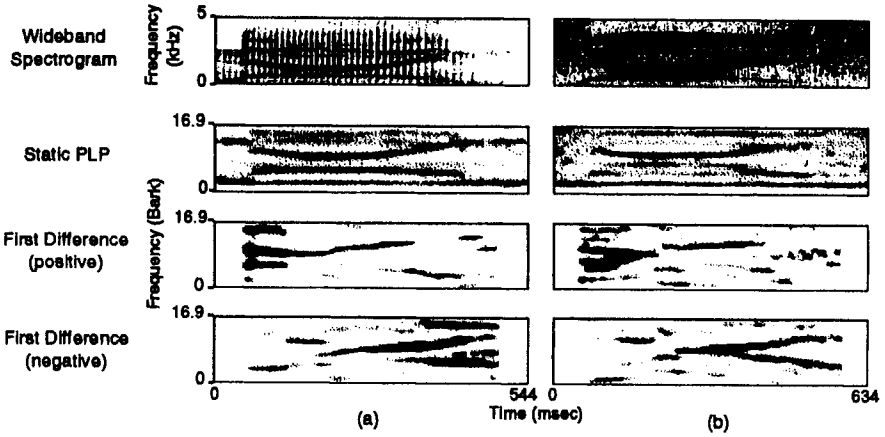

Figure 1 Standard wideband spectrogram, and static, first difference positive component, and first difference negative component pseudo-spectrograms for: a) clean speech and b) noisy Lombard speech from the same talker (see [24, 38] for further details).

static spectrograms appear to have been removed in the difference representations. Also note that there is some consistency apparent across different production conditions (i.e. the normal speech of Fig. 1a versus the Lombard speech in noise of Fig. 1b).

3.2 Temporal Derivative Features

3.2.1 Background and Definitions

As described above, one way to utilize dynamic features of speech spectra is through approximation of the temporal derivative as implemented by the first difference of eqn. 1. Early automatic speech recognition experiments utilizing first difference dynamic features are described in [39–42]. In various speech recognition experiments [24, 25, 43, 44], further improvements have also been obtained utilizing the *second difference*, as defined below:

$$A(f, n\Delta T) = D(n\Delta T + \delta_A) - D(n\Delta T - \delta_A) \qquad (2)$$

Note the time delay interval used here, δ_A, is not necessarily the same as that in the first difference. This raises the issue of the temporal extent of the data used to approximate the derivatives. Although differences can be taken between adjacent frames, most researchers calculate derivatives over longer intervals, based on the assumption that the time scale of useful dynamic information in

speech is of longer extent than the typical 10 to 20 msec frame step size of most speech analyses.

The intervals over which the derivatives are calculated will be referred to as the *derivative window*, and should not be confused with the time over which the original analysis is calculated (i.e. the *analysis window*). For the first and second difference cases of eqns. 1 and 2, the derivative windows are $2\delta_D$ and $2(\delta_D + \delta_A)$, respectively.

As the difference implementation estimates the time derivative of the spectral feature using the least number of speech analysis frames, it can introduce significant error. If a smoother estimate of the derivative is required, numerical regression approaches can be applied [45]; these typically use all of the speech frames in the derivative window. The regression formula for approximating the rth-order derivative of spectral parameter $S(t)$ is:

$$R_r(t, K, \Delta T) = \frac{\sum_{k=-(K+1)/2}^{(K+1)/2} P_r(k, K) S[t + k\Delta T]}{\sum_{k=-(K+1)/2}^{(K+1)/2} P_r^2(k, K)} \tag{3}$$

where ΔT is the frame step size and the weighting function $P_r(k, K)$ is the rth-order polynomial to be applied to K frames (K an odd integer). The first few orthogonal polynomials [45] are:

$$P_0(k, K) = 1$$
$$P_1(k, K) = k$$
$$P_2(k, K) = k^2 - \frac{1}{12}(K^2 - 1)$$

For simplicity, the time dependence will be dropped, e.g. the regression approximation to the first derivative over a 40 msec window will be denoted by $R_1(40)$. Note that regression window length is defined here as the interval between the centers of the first and last frames.

Recognition experiments with regression-implemented first derivative features, done by Furui [20] and Soong *et al.* [21], reported impressive recognition improvements, which were the impetus for much further investigation of dynamic features. Subsequent work has demonstrated significant, though lesser, gains when second order regressions (alternatively implemented as a difference of first order regressions) were applied to various recognition problems including noisy/degraded isolated words [43, 46], large vocabulary continuous speech [47], and continuous digits [48]. In [47] it was shown that the addition of the

second derivative coefficients to the speech parametric representation is not always beneficial for every speaker, even if the overall performance is improved. Even third derivatives have been found to give small, but significant recognition gains [46] in recognition of a confusable vocabulary. However, these gains were not found on a less confusable vocabulary [43]. Along with second derivative, Huang *et al* [44] also incorporated hierarchical features (representing short-term and long-term spectral dynamics [23]) in their continuous speech recognition system. They reported a 15% error reduction as compared to the use of static and first derivative.

From the above it can be seen that derivative features have seen wide application in speech recognition. Recent studies have shown that they are particularly useful for dealing with different types of adverse recognition conditions. In the next section we will review several such experiments.

3.2.2 Isolated Word Recognition using Temporal Derivatives

This section considers various issues in the application of derivative features for speaker-independent isolated word recognition in the presence of additive noise and/or Lombard effects. Application of spectral dynamics in continuous speech recognition is considered in Section 3.5.

Recognition system and databases

The recognition front-end used in these experiments produces cepstral coefficients from PLP analysis [34, 35]. PLP combines autoregressive modeling with three psychophysically motivated speech transformations: 1) Bark-space filterbank processing, 2) equal-loudness compensation, and 3) third root intensity-loudness compression. The derivative features are calculated from the cepstral coefficients using the regression approximation of eqn. 3, previous work having shown that the regression implementation of derivatives consistently outperforms the difference implementation [43, 49]. The isolated word recognizer is based on whole word, discrete-density hidden Markov models.[2]

An isolated word database of 21 confusable words was used for evaluation. The words were English alpha-digits and control words, comprising five confusable subsets: "a j k", "b c d e g p t v z three", "m n", "go no oh", and "f s x". The speech data were recorded under both noise-free (clean) and noisy (Lombard) conditions. In the case of Lombard speech, talkers were exposed to 85 dB

[2] With the exception of the experiment summarized in Fig. 4, all of the isolated word recognition results presented in this chapter come from the recognition system and "Confusable Words" database described above (see [24, 43] for more details).

SPL noise through headphones. White noise at 18 dB SNR was later added to the speech data to simulate noisy speech conditions. The data were then divided into disjoint sets for training and testing, with clean data always used for training.

Recognition results

The results shown in Fig. 2 were obtained by evaluating the speaker-independent recognition rates from front-ends incorporating different combinations of static and derivative features. Various conclusions can be drawn from these results:

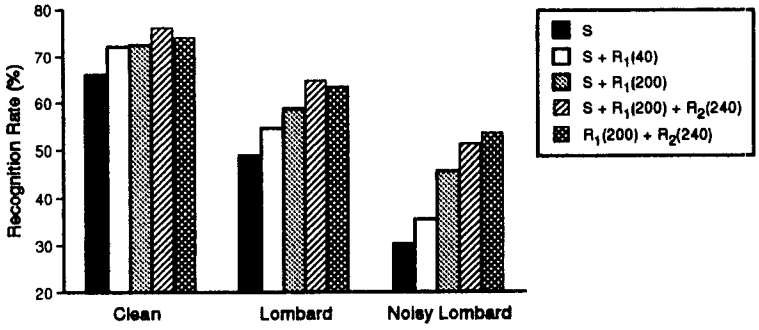

Figure 2 Recognition results showing improvements in various conditions using temporal derivative features (adapted from [50]).

Improvements from first derivative: Overall the biggest improvements in recognition rate come from supplementing the static feature (S) with the first derivative (approximated here by the regression R_1). As a result, it is obvious why nearly all recognizers currently use some form of first derivative spectral feature in their speech representation

Effects of derivative window lengths: In Fig. 2, "S + R_1(40)" represents the static feature (S) combined with a 40 msec regression-evaluated derivative; similar short-window length derivative features have been implemented in many previous recognition systems (e.g. [40, 41]). The other cases of Fig. 2 utilize a much longer, 200 msec R_1 window, which was chosen as a result of the experiments summarized in Fig. 3 (from [46]). Here the window lengths were varied for tests with clean and noisy Lombard speech data. Although longer first derivative windows do not give any significant gains for clean data (suggesting why many researchers use short windows), there are large gains with longer first derivative windows for the noisy-Lombard data. Second derivatives based on longer windows are useful for both the clean and noisy-Lombard data.

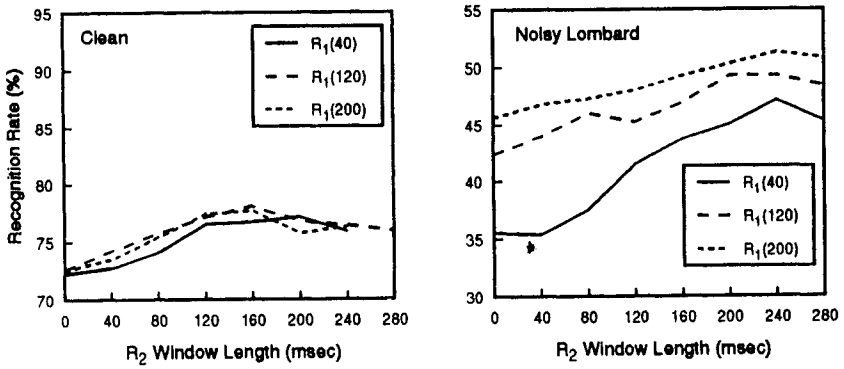

Figure 3 Recognition rate versus R_2 window length for the combination of static, R_1, and R_2 features. Test speech was clean (left graph) or noisy Lombard at 18 dB SNR (right graph).

Similar advantages of longer derivative windows have been demonstrated in other isolated word recognition studies, for example with digit [24] and multi-syllable Japanese city name databases [49]. However, in all of these cases whole word units were used for recognition, so "spreading" of spectral transitions across phoneme boundaries, induced by the use of long-window derivatives, was apparently not a problem (even for the Confusable Words data used in Fig. 3). Such "transition spreading" could be more of a problem in continuous or subword-unit-based speech recognition. Use of long derivative windows in continuous speech recognition is considered further in Section 3.5.1.

Mismatched training and test conditions: As mentioned earlier, the experiments considered in this section used clean training data. It is apparent from Figs. 2 and 3 that derivative features (particularly long windows and second derivatives) help most in the noisy Lombard cases, where the mismatch of test and training conditions is greatest.

To investigate this further, recognition experiments were run where the speech representation consisted of only two dynamic features; these are shown in Fig. 2 by the cases labeled "$R_1 + R_2$". Although dropping the static feature decreases the recognition rate for matched conditions, it increases recognition rate in the noisy Lombard case; this suggests that dynamic features are more robust than the static feature in highly mismatched conditions. Other front-end analyses, which also emphasize the dynamic over the static parts of the spectral representation (although less explicitly than by deleting the static feature), have been shown to be similarly robust. These analysis methods will be discussed in the next two sections.

3.3 Spectral Mean Normalization

The preceding section reviewed techniques using spectral dynamics in speech recognition through direct application of temporal derivative features. Other techniques implicitly utilize spectral dynamics by suppressing the slower varying parts of the spectral representation, based on the assumption that the slower variations are generally due to channel effects and carry little phonetic information.

Assuming the channel distortion can be represented by a linear transfer function and that the noise is negligible, the spectrum of the input to the recognizer, $Y(f, n\Delta T)$, is the product of the original speech spectrum, $S(f, n\Delta T)$, and the channel transfer function $|H(f, n\Delta T)|^2$. Attempts to remove channel distortions are then often done in a domain which is linear with respect to the log power spectrum, where the distortion is additive:

$$\log Y(f, n\Delta T) = \log S(f, n\Delta T) + \log |H(f, n\Delta T)|^2 \quad (4)$$

Suitable domains include log spectra, log filterbank powers, log all-pole model spectra, and cepstra (which are linear transforms of the log power spectra). Note that with all-pole model spectra the analysis order must be sufficient for accurate modeling.

Based on eqn. 4, suppressing channel distortions may be viewed as a spectral normalization or a parameter filtering operation. *Spectral parameter filtering* to suppress $|H(f, n\Delta T)|^2$, while preserving the phonetically relevant parts of $S(f, n\Delta T)$, is the topic of the next section. *Spectral normalization*, or subtracting from $\log Y(f, n\Delta T)$ an estimated spectral bias (which includes $\log |H(f, n\Delta T)|^2$ as a component), is discussed below.

Spectral normalization requires an estimated bias vector for each analysis frame. A variety of techniques have been proposed which make use of *a priori* information about the speech and channel distortion to determine the bias. Estimating the bias vector from the input data *alone* requires further assumptions. For example, the distortion may be assumed to be stationary, so that a simple averaging operation will suffice. Alternatively, it can be assumed to be constant conditioned on the vector quantization codeword [51], phone label [15], hidden Markov model state [52, 53], or decoded word string [16].

In a number of studies (e.g. [10, 54, 55]) the average of the input speech spectrum over the whole utterance has been used as the bias vector. Although subtracting this long-term estimate of the bias suppresses some speech characteristics, only the most slowly varying components of speech are affected and

speech phonetic information is preserved. With this averaging approach the channel effects can be suppressed as follows:

$$\begin{aligned}\log \hat{S}(f, n\Delta T) &= \log Y(f, n\Delta T) - \log |\hat{H}(f, n\Delta T)|^2 \\ &\approx \log Y(f, n\Delta T) - \sum_{\text{utterance}} \log Y(f, n\Delta T)\end{aligned} \tag{5}$$

As subtraction of log spectra is equivalent to division of power spectra, such an approach is commonly referred to as "spectral normalization."

The initial applications of this normalization approach were for speaker verification problems [54, 55]. As in many current recognizers, these earlier works dealt with speech parameterization in the cepstral domain (i.e. cepstral coefficients were used in eqn. 5), so the technique was referred to as *cepstral mean normalization*. Since the initial applications of cepstral mean normalization, many researchers have verified that this simple technique provides significant speech recognition rate increases in various conditions. Accordingly, cepstral mean normalization has become a standard technique in many recognition systems (e.g. [10, 13, 56, 57]).

One example of the effectiveness of cepstral mean normalization is given by Liu *et al.* [10], where it was applied for compensation of mismatches from a wide-range of testing microphones (i.e. desktop, stand-mounted, telephone hand-sets, speaker phones, etc.). Liu *et al.* found in tests with continuous speech recognition (ARPA 5000-word "Wall Street Journal" task) that the average word error rate with the various microphones was reduced from 38.5% to 21.4% by simply applying cepstral mean normalization. This work also considered additional techniques (requiring training data simultaneously recorded from different kinds of microphones) to be used in combination with cepstral mean normalization to get further improvements; but the fact that cepstral mean normalization alone reduced the "mismatched-microphone" error rate by almost half shows the utility of this simple approach.

One drawback of the non-causal cepstral mean normalization of eqn. 5 is that the average is calculated over the whole utterance. In many recognition applications the resulting delays are unacceptable. One solution is to estimate the spectral average over the preceding signal:

$$\log \hat{S}(f, n\Delta T) = \log Y(f, n\Delta T) - \sum_{k}^{K} a_k \log Y[f, (n-k)\Delta T] \tag{6}$$

Making cepstral mean normalization causal, as in eqn. 6, has been shown to work well using either long-term averages [56] or short-term averages [58, 59] for the estimated cepstral mean.

Note the moving average of eqn. 5 has been replaced by a weighted average, so that eqn. 6 has the form of a general finite impulse response (FIR) filter. This emphasizes that cepstral mean normalization can also be interpreted as a filtering operation.

3.4 Spectral Parameter Filtering

The spectral parameters input to the speech recognizer, $Y(f, n\Delta T)$, can be viewed as a time series that represents the temporal variations of the speech spectra. This time series will itself have spectral components at frequencies up to half of the analysis frame rate (e.g. 50 Hz for a 10 msec frame step). For example, if the speech is analyzed into subband parameters, the spectral components will occur at frequencies centered around the subband envelope variation rate, i.e. the "modulation frequency" [60, 61]. In this chapter, the term modulation frequency will be used in a general sense to describe the rate of change of any spectral parameter representation.

Spectral mean normalization required the assumption that the input to the recognition system is the sum of the speech spectrum and the channel transfer function in a domain which is logarithmic with respect to power spectra, as in eqn. 4. With the further assumption that the channel effects and the (phonetically significant parts of the) speech occupy different modulation frequencies, channel effects can be suppressed by linear filtering.

In [60] it was shown that the energy of speech in a subband representation primarily occurs within a narrow band of modulation frequencies, e.g. 1 to 8 Hz for connected discourse filtered by an octave band-pass filter centered at 1 kHz. The peak modulation frequency for this case was found to be at about 3 Hz, which corresponded to the syllable rate. Houtgast *et al.* go on to suggest that the relevant range of subband modulation frequencies for intelligible speech reproduction is approximately 0.4 to 20 Hz. Thus channel variations which occur at modulation frequencies less than this speech frequency range can be suppressed by high-pass filtering (e.g. with a pass-band starting at 0.4 Hz).

Both of the techniques discussed so far (temporal derivatives and spectral normalization) can be interpreted in terms of spectral parameter filtering. That is, they are equivalent to a (FIR) filtering that rejects lower modulation frequency variations of the speech parameters.

3.4.1 Subband Parameter Filtering

Papers by Hermansky *et al.* [26] and Hirsch *et al.* [27] introduced an alternative approach to spectral mean normalization. This approach is based on suppressing the lower modulation frequencies of a subband analysis using a band-pass characteristic realized by infinite impulse response (IIR) filtering, e.g. [14]:

$$W(z) = \frac{0.2z^4 + 0.1z^3 - 0.1z^1 - 0.2}{1 - \rho z^{-1}} \tag{7}$$

The position of the spectral pole (ρ), usually chosen empirically, determines the time-constant of the parameter filter (e.g. 160 msec for $\rho = 0.94$).

Hermansky *et al.* proposed incorporating such IIR filtering into PLP (spectral) analysis, referring to this as the *RelAtive SpecTrAl*, or "RASTA", technique. The term RASTA has since been applied by numerous researchers in describing a variety of similar techniques [10, 12, 62, 63]. However, to avoid confusion, use of this term will be restricted here to its original implementation in the "RASTA-PLP" algorithm.

Parameter filtering is incorporated into PLP analysis, yielding RASTA-PLP [26], by adding three steps to the standard PLP algorithm. These steps, inserted after the first stage of PLP (see Section 3.2.2), are:

1.1 Take logs of the filter bank energies.
1.2 Filter these log energies utilizing eqn. 7.
1.3 Take anti-logs of the filtered result.

Since the spectral parameters to be filtered here are filterbank energies, RASTA-PLP is closely related to Hirsch's subband filtering technique [27]. Accordingly experiments using RASTA-PLP will be described below.

The relative advantages of band-pass filtering of subband energies (i.e. as implemented in RASTA-PLP) are examined for various experimental conditions in [14]. An interesting example from this work considers speaker-independent isolated digit (HMM-based) recognition under adverse conditions, i.e. test speech with additive noise and/or a constant linear distortion. Recognition performance is evaluated for front-ends consisting of PLP alone, PLP with cepstral mean normalization, and RASTA-PLP. Additionally, results are given for ideal "matched-condition training" cases, where the recognizer is trained under the same conditions as the test data. In all other cases the recognizer is trained on clean speech, regardless of the test speech condition. As indicated by Fig. 4, RASTA-PLP performs very well where expected, i.e. when the linear distortion alone is present. For this case, RASTA-PLP (and also PLP with cepstral

mean normalization) improves the recognition rate to nearly its original (clean, undistorted data) value. However, RASTA-PLP does not perform well when applied to noisy data (such problems being inherent to spectral parameter filtering/normalization techniques in general). Proposals to handle such additive noise problems will be discussed in Section 3.4.3.

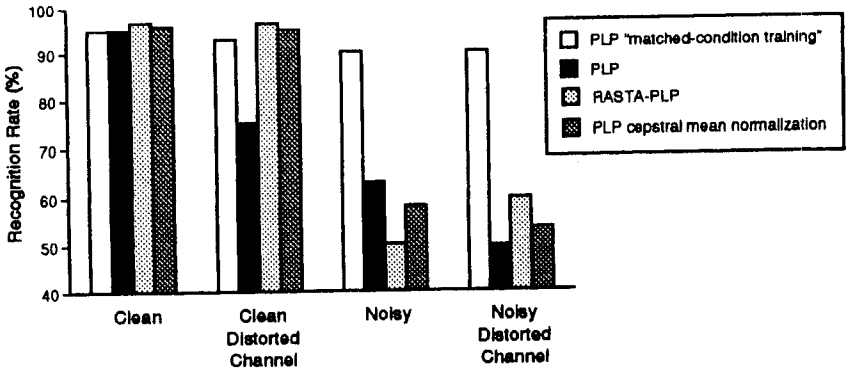

Figure 4 Comparison of digit recognition in noisy and/or distorted channel conditions for front-ends consisting of PLP alone, PLP with cepstral mean normalization and RASTA-PLP (adapted from [14]).

3.4.2 Other Spectral Parameter Filtering Techniques

Spectral parameter domains

In addition to considering spectral parameter filtering of log subband energies, more recent work has considered recognition based on parameter filtering of other log spectral parametric representations. Recognition improvements have been demonstrated under various adverse conditions from application of parameter filtering in LP cepstral [63], mel-cepstral [62], or PLP cepstral [58] domains. The last case, filtering of PLP cepstral coefficients (which differs from RASTA-PLP in that parameter filtering is done *after* PLP analysis) was shown to give results similar to RASTA-PLP.

Spectral parameter filter implementations

The form of the spectral parameter filter has received further examination. Band-pass (IIR, eqn. 7) and high-pass (FIR, eqn. 6) filter implementations were found to give similar results for isolated word recognition in [58]. Further improvements have been obtained by adjusting the spectral parameter filters

according to parameter index. In the log subband energy domain, Smolders and Van Compernolle [63] demonstrated improved recognition rate by making the spectral parameter filter time constants shorter for higher critical bands. In the cepstral domain, Aikawa *et al.* [29] shortened the impulse response of the spectral parameter filters for higher cepstral coefficient indices, based on psychophysical considerations.

3.4.3 Enhancements to Spectral Parameter Filtering

Approaches to handling additive noise

Spectral normalization and parameter filtering assume that channel distortions are linear *and* that additive noise is negligible. If the noise is not negligible, then the spectral input to the recognizer contains an additional term, i.e.:

$$Y(f, n\Delta T) = S(f, n\Delta T)|H(f, n\Delta T)|^2 + N(f, n\Delta T) \quad (8)$$

As the noise and channel distortions are additive in different domains (i.e. the log and power domains respectively), they cannot be simultaneously suppressed by spectral parameter filtering as described above. One approach to this problem is to sequentially transform between domains to suppress noise and channel distortions in the linear and log domains respectively. Preprocessing techniques for noisy speech include spectral subtraction [64] and spectral mapping [65]. An example of this sequential approach, using spectral subtraction followed by RASTA-PLP, is seen in [11]. Others addressed both noise and channel distortion simultaneously by compensating the HMM output parameters for the effects of mismatch between test and reference data during the pattern matching stage of recognition (e.g. [16, 66]).

Hermansky and Morgan [14] proposed a modified spectral parameter filtering technique for handling additive noise. Their "Lin-log RASTA" approximates both spectral subtraction (when the signal power is low relative to noise) and the removal of channel distortions (when the signal power is high). This was achieved by replacing the log transformation in RASTA-PLP (step 1.1 in Section 3.4.1) with a function which is nearly linear for low values and logarithmic for high values. The inverse relation of step 1.3 is also appropriately modified. As the algorithm depends upon relative signal and noise levels, the transformation is adapted according to an estimate of the instantaneous signal-to-noise ratio. When their best adaptive approach was compared to the standard RASTA-PLP results shown in Fig. 4, large recognition rate increases were found for the noisy conditions, with only a small loss of recognition rate in the clean distorted-channel condition.

Derivatives of filtered spectral parameter features

As described in Section 3.2, augmenting static spectral features with their first time derivative greatly improves speech recognition performance, particularly under adverse conditions; however, second and higher derivatives do not always provide improvements. Since features from spectral parameter filtering inherently incorporate some form of derivative (e.g. see the numerator of eqn. 7), there was doubt about whether these features would be improved by augmenting them with their derivatives. However, as demonstrated in [58], significant recognition improvements can be obtained when spectral parameter filtered features are augmented with their derivatives.

Examples from [58] are shown in Fig. 5 for experiments on the Confusable Words database using "clean" speech, "distorted channel" speech (i.e. passed through a fixed, second-order pole, band-pass filter [62]), and Lombard speech with additive noise at 18 dB SNR. Recognition rates are given for the static feature alone and static combined with derivative features, where the first and second derivatives are evaluated by regressions $R_1(80)$ and $R_2(240)$, respectively. As expected, improvements are obtained in all cases when the static PLP feature is augmented with derivatives. More importantly, utilizing derivatives of the spectral parameter filtered feature (RASTA-PLP) also gives significant improvement in all cases. Even the second regression of RASTA-PLP, which resembles a third derivative of static PLP, gives improvements.

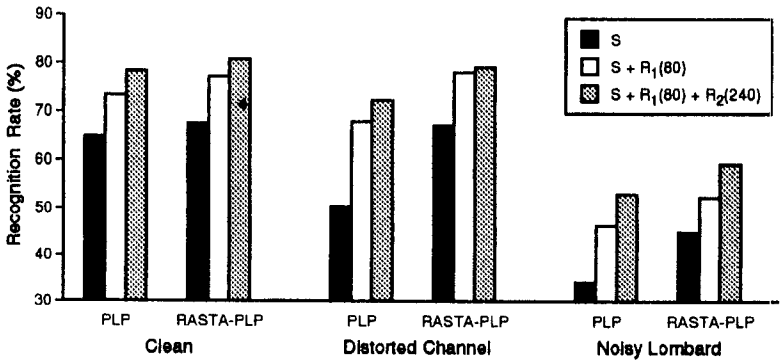

Figure 5 Comparison of recognition rates based on PLP and RASTA-PLP analyses augmented with their temporal derivatives.

3.5 Spectral Dynamics for Continuous Speech Recognition (CSR)

This section compares results from speaker-independent continuous speech recognition with some of the key results, presented in previous sections, for isolated word recognition.

3.5.1 Regression Features Applied to CSR

For continuous speech recognition, as for isolated word recognition, use of the first derivative yields large improvements (e.g. [41]). The introduction of the second derivative yields additional improvement in recognition rates, although the amount of improvement is less than that obtained with the first derivative. A number of studies have presented data in agreement with these conclusions (e.g. [25, 44, 67]).

Various insights are provided by a recent study [68] of the performance of a multi-pass, HMM-based recognizer when evaluated on a telephone speech database [69] for a spelled-name task. In this study, continuous-letter recognition was evaluated for both clean and distorted channel telephone speech (the distorted channel condition being the same as in the experiments summarized in Fig. 5). The pertinent conclusions from [68] are as follows:

1. Incorporating a second derivative feature (implemented with R_2) gives a large improvement in recognition accuracy in the mismatched distorted channel case, and only slightly improves recognition for undistorted speech.
2. As with isolated word recognition, the combination of first and second derivatives without the static feature (i.e. $R_1 + R_2$) compensates well for cases with significant mismatch between training and testing conditions. However, $R_1 + R_2$ alone decreases the recognition accuracy for clean (i.e. undistorted) test data.
3. Long regression windows for the first and second derivatives decrease recognition accuracy (the average letter duration here is 386 msec). Additional experiments for other window sizes confirmed this observation.

The fact that short regression windows work better than long regression windows for continuous speech, and conversely that long windows are more suitable for isolated words, has also been recently noted by Nadeu and Juang [70].

3.5.2 Spectral Parameter Filtering Applied to CSR

Depending on the speech conditions, the effectiveness of spectral parameter filtering for isolated word recognition has been shown to be quite variable. For example, it provides large improvements when linear channel distortions are present but little improvement (and sometimes small losses [14, 58]) for clean, undistorted speech. This issue is re-examined for continuous speech recognition by comparing the results of PLP and RASTA-PLP shown in Fig. 6. These results are from the same recognizer and database described in the previous section (see [68] for details).

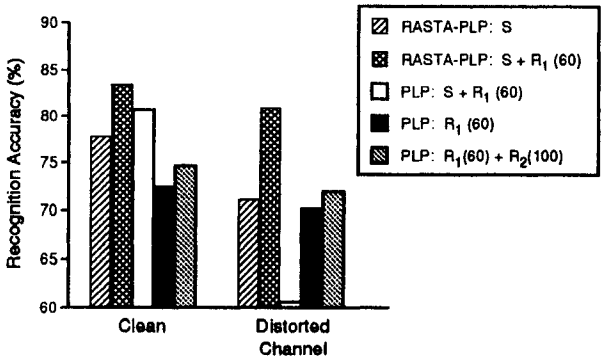

Figure 6 Comparison of letter recognition accuracy for RASTA-PLP and PLP based feature sets in continuous spelled-name recognition.

As shown in Fig. 6 (by comparing the "S + R_1" cases of RASTA-PLP and PLP), the spectral parameter filtering in RASTA-PLP increases recognition accuracy for the clean and distorted channel cases by approximately 3% and 20% respectively. Thus, in agreement with earlier work (e.g. [10, 14, 62]), spectral parameter filtering is shown to be useful for continuous speech recognition. Another point evident from Fig. 6 is that significant additional recognition gains can be obtained by augmenting the spectral parameter filtered feature (RASTA-PLP here) with its first derivative, in agreement with conclusions noted earlier for isolated words.

It has already been shown that spectral parameter filtering is *not* suitable for continuous speech recognition based on context-independent subword units [13, 14]. Typical parameter filters involve an impulse response that is too long for context-independent subword units. However, as noted in [14], parameter filtering works well with whole word models (such as letter models, as shown above) or in phoneme-based recognizers that use broad temporal input contexts, such as triphones. This result is analogous to the conclusion of Section

3.5.1, i.e. the optimal time window length of derivative features is shorter for continuous speech than for isolated words.

Finally, temporal derivative features and the spectral parameter filters considered here have similar computational forms, and both techniques suppress lower modulation frequencies. To compare these techniques in terms of recognition performance, results for derivative features alone are included in Fig. 6. Note that RASTA-PLP yields generally better results than those obtained from the first derivative (R_1) of PLP. Additionally, RASTA-PLP combined with its first derivative considerably outperforms the combination of $R_1 + R_2$ on both clean and distorted channel speech. These results indicate that the spectral parameter filtering in RASTA-PLP processing is more than simply a first derivative effect, i.e. the temporal integration provided by the denominator term in eqn. 7 also contributes to recognition performance. Various explanations for these results and ways to realize "optimal" dynamic features have been proposed [14, 58, 70]; these will be discussed in the next section.

4 DISCUSSION

Much of the phonetic information in speech is encoded in the changes of the speech spectrum over time (see e.g. [32, 39, 71]). Temporal derivative features and spectral parameter filtering exploit, in different ways, the information contained in the changes of the speech spectrum. Such techniques have been shown to improve recognition performance under adverse conditions such as additive noise, channel distortions, or Lombard effect. So it might be asked: *How do techniques that exploit spectral dynamic information improve speech recognition performance?* This question may be addressed from several perspectives.

In hidden Markov model based recognition systems, dynamic features partially compensate for the fact that, at typical frame rates, speech strongly violates the assumption of independent observations. Including speech spectral dynamics in the HMM observation parameters models the amount of spectral change between successive frames in a manner that is consistent with allowable speech productions. However, this interpretation does not explain the advantages found for using spectral dynamics with template-based recognition techniques, such as dynamic time warping.

Further insights are provided by interpreting a sequence of dynamics-based features as the output of a linear filter applied to the time series of static spectral parameters (i.e. as discussed in Section 3.4). Time derivative features and the kinds of spectral parameter filters considered in this chapter share two basic components: a differentiation and a temporal smoothing. Nadeu

and Juang [70] argue that the *temporal smoothing* bounds the pass-band to lower modulation frequencies, which may then be more reliably estimated, and that the differentiation *equalizes* the parameter power spectrum (i.e. the power spectrum of the time series composed of spectral parameters).

The degree of temporal smoothing from a parameter filter is strongly influenced by the length of its time response. The optimal amount of temporal smoothing depends on the upper bound of the phonetically relevant frequencies in the parameter power spectrum. In [70], long-term average parameter power spectra are derived from LP cepstral coefficients by averaging over cepstral indices and many utterances. Comparison of these spectra, for DARPA Resource Management and isolated digit data, showed a broader parameter power spectrum for continuous speech than for isolated word data. This is consistent with our observation that long-window derivatives are more suitable for recognition of isolated words than for continuous speech.

Equalization of the parameter power spectrum can be viewed as decorrelating the sequence of spectral parameters, thereby enhancing temporal resolution and satisfying the "independent observations" assumption of HMM-based recognition systems. The shape of the average spectrum of cepstral coefficients examined in [70] is well approximated by a first-order pole near the phonetically important low modulation frequencies, and can therefore be equalized by a matching zero in a differentiation filter. Detailed examination of parameter power spectral characteristics for individual parameters in different parameter domains promises new insight into the speech recognition performance tradeoffs between the time resolution and spectral equalization of feature parameters.

The view of dynamic features as filtering in a parameter power spectral domain also helps explain the complex interactions observed between filtered parameter features when they are combined in a speech representation. In particular, good recognition results are expected when the pass-bands of different features adequately cover the phonetically relevant portion of the parameter power spectrum. Complex interactions between long-window derivative features may arise from the need to cover modulation frequencies suppressed by one feature with a pass-band of another feature. Finally, as feature sets become more complex (e.g. by combining filtered features with their higher temporal derivatives) the interaction between the features increases. Techniques such as linear discriminant analysis [72] may help deal with this interaction.

Spectral dynamics have been successfully applied in speech recognition through a wide variety of techniques; a few of the simpler of these techniques have been discussed in this chapter. By exploiting the basic concept that "only the changes bear information" [73], considerable progress has been obtained in the

short time since temporal derivatives were introduced in speech recognition. However, automatic speech recognition performance still falls far short of human capabilities, especially in the presence of irrelevant talker, environmental, or channel variations. More study of the basic attributes of speech, including spectral dynamics, is required to find a compact and robust speech representation which can extract phonetically relevant information under adverse speech recognition conditions.

REFERENCES

[1] J. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, pp. 1215–1247, Sept. 1993.

[2] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *JASA*, vol. 84, pp. 917–928, 1988.

[3] J. Hansen, *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition.* PhD. thesis, Georgia Institute of Technology, 1988.

[4] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *JASA*, pp. 510–524, 1993.

[5] J. Pickett, "Effects of vocal force on the intelligibility of speech sounds," *JASA*, vol. 28, pp. 902–905, 1956.

[6] J. Dreher and J. O'Neill, "Effects of ambient noise on speaker intelligibility for words and phrases," *JASA*, vol. 29, pp. 1320–1323, 1957.

[7] F. Soong and M. M. Sondhi, "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise," *IEEE Trans. ASSP*, vol. 36, no. 1, pp. 41–48, 1988.

[8] D. Mansour and B.-H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. ASSP*, vol. 37, no. 11, pp. 1659–1671, 1989.

[9] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition.* PhD thesis, Carnegie Mellon University, 1990.

[10] F.-H. Liu, R. Stern, A. Acero, and P. J. Moreno, "Environment normalization for robust speech recognition using direct cepstral comparison," *Proc. ICASSP*, vol. II, pp. 61–64, 1994.

[11] J. Smolders, T. Clase, G. Sablon, and D. VanCompernolle, "On the importance of the microphone position for speech recognition in the car," *Proc. ICASSP*, vol. I, pp. 429–432, 1994.

[12] J. Chang and V. Zue, "A study of speech recognition system robustness to microphone variations: Experiments in phonetic classification," *Proc. ICSLP*, vol. 3, pp. 995–998, 1994.

[13] H. Van hamme, G. Gallopyn, L. Weynants, B. D'hoore, and H. Bourlard, "Comparison of acoustic features and robustness tests of a real-time recognizer using hardware telephone line simulator," *Proc. ICSLP*, pp. 1907–1910, 1994.

[14] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[15] Y. Zhao, "Iterative self-learning speaker and channel adaptation under various initial conditions," *Proc. ICASSP*, vol. 1, pp. 712–715, 1995.

[16] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," accepted for publication in *IEEE Trans. Speech and Audio Processing*.

[17] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, April 1995.

[18] S. Furui, "Toward robust speech recognition under adverse conditions," *Proc. ESCA Workshop on Speech Processing in Adverse Conditions*, pp. 31–42, Nov. 1992.

[19] B.-H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, vol. 5, pp. 275–294, 1991.

[20] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. ASSP*, vol. 34, pp. 52–59, 1986.

[21] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *Proc. ICASSP*, pp. 877–880, 1986.

[22] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," *Proc. ICASSP*, pp. 1991–1994, 1986.

[23] S. Furui, "On the use of hierarchical spectral dynamics in speech recognition," *Proc. ICASSP*, pp. 789–792, 1990.

[24] B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech," *Proc. ICASSP*, pp. 857–860, 1990.

[25] H. Ney, "Experiments on mixture-density phoneme-modelling for the speaker-independent 1000-word speech recognition task," *Proc. ICASSP*, pp. 713–716, 1990.

[26] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," *Proc. EUROSPEECH*, pp. 1367–1370, 1991.

[27] H. G. Hirsch, P. Meyer, and H. W. Ruehl, "Improved speech recognition using high-pass filtering of subband envelopes," *Proc. EUROSPEECH*, pp. 413–416, 1991.

[28] T. Kitamura, E. Hayahara, and Y. Simazaki, "Speaker-independent word recognition in noisy environments using dynamic and averaged spectral features based on a two-dimensional mel-cepstrum," *Proc. ICSLP*, pp. 1129–1132, 1990.

[29] K. Aikawa, H. Singer, H. Kawahara, and Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition," *Proc. ICASSP*, vol. II, pp. 668–671, 1993.

[30] B. P. Milner and S. V. Vaseghi, "Speech modeling using cepstral-time feature vectors," *Proc. ICASSP*, vol. 1, pp. 601–604, 1994.

[31] H.-F. Pai and H.-C. Wang, "A study of the two-dimensional cepstrum approach for speech recognition," *Computer Speech and Language*, vol. 6, pp. 361–375, 1992.

[32] S. Furui, "On the role of spectral transition for speech perception," *JASA*, pp. 1016–1025, 1986.

[33] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*. Springer-Verlag, 1976.

[34] H. Hermansky, B. Hanson, and H. Wakita, "Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain," *Speech Communication*, vol. 4, pp. 181–187, 1985.

[35] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *JASA*, vol. 87, no. 4, pp. 1738–1752, 1990.

[36] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, vol. 28, pp. 357–366, Aug. 1980.

[37] B. A. Hanson and H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise," *IEEE Trans. ASSP*, vol. 35, pp. 968–973, 1987.

[38] T. H. Applebaum and B. A. Hanson, "Perceptually-based dynamic spectrograms," in *Visual Representations of Speech Signals*, edited by M. Cooke, S. Beet, and M. Crawford, ch. 11, pp. 153–160, Wiley, 1993.

[39] K. Elenius and M. Blomberg, "Effects of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system," *Proc. ICASSP*, pp. 535–538, 1982.

[40] V. N. Gupta, M. Lennig, and P. Mermelstein, "Integration of acoustic information in a large vocabulary word recognizer," *Proc. ICASSP*, pp. 697–700, 1987.

[41] K.-F. Lee, *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*. PhD thesis, Comp. Sci. Dept., Carnegie Mellon University, 1988.

[42] K. Shikano, "Evaluation of LPC spectral matching measures for phonetic unit recognition," CMU-CS-86-108, Comp. Sci. Dept., Carnegie Mellon University, 1986.

[43] T. H. Applebaum and B. A. Hanson, "Robust speaker-independent word recognition using spectral smoothing and temporal derivatives," *Signal Processing V - Proc. EUSIPCO*, pp. 1183–1186, Elsevier Science, 1990.

[44] X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, K.-F. Lee, and R. Rosenfeld, "The SPHINX-II speech recognition system: An overview," *Computer Speech and Language*, vol. 2, pp. 137–148, 1993.

[45] N. R. Draper and H. Smith, *Applied Regression Analysis*. New York: Wiley, 1981.

[46] T. H. Applebaum and B. A. Hanson, "Features for speaker-independent recognition of noisy and Lombard speech," *JASA Suppl. 1*, vol. 88, Fall 1990. Reprinted in *J. of Amer. Voice I/O Soc.*, vol. 14, pp. 73-80, 1993.

[47] C.-H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini, and A. E. Rosenberg, "Improved acoustic modeling for continuous speech recognition," *Proc. DARPA Workshop on Speech Recognition*, pp. 319–326, DARPA, 1990.

[48] J. G. Wilpon, C.-H. Lee, and L. R. Rabiner, "Connected digit recognition based on improved acoustic resolution," *Computer Speech and Language*, vol. 7, pp. 15–26, 1993.

[49] T. H. Applebaum and B. A. Hanson, "Tradeoffs in the design of regression features for word recognition," *Proc. EUROSPEECH*, pp. 1203–1206, 1991.

[50] B. A. Hanson and T. H. Applebaum, "Features for noise-robust speaker-independent word recognition," *Proc. ICSLP*, pp. 1117–1120, 1990.

[51] A. Acero and R. M. Stern, "Robust speech recognition by normalization of the acoustic space," *Proc. ICASSP*, pp. 893–896, 1991.

[52] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. ASSP*, vol. 37, pp. 1846–1856, 1989.

[53] V. L. Beattie and S. J. Young, "Noisy speech recognition using hidden Markov model state based filtering," *Proc. ICASSP*, pp. 917–920, 1991.

[54] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *JASA*, vol. 55, pp. 1304–1312, 1974.

[55] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. ASSP*, vol. 29, pp. 342–350, 1981.

[56] D. Geller, R. Haeb-Umbach, and H. Ney, "Improvements in speech recognition for voice dialing in the car environment," *Proc. ESCA*

Workshop on Speech Processing in Adverse Conditions, pp. 203–206, Nov. 1992.

[57] R. Schwartz, T. Anastasakos, F. Kubala, J. Makhoul, L. Nguyen, and G. Zavaliagkos, "Comparitive experiments on large vocabulary speech recongition," *Proc. ARPA Workshop on Human Language Tech.*, March 1993.

[58] B. A. Hanson and T. H. Applebaum, "Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech," *Proc. ICASSP*, vol. II, pp. 79–82, 1993.

[59] A. E. Rosenberg, C.-H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," *Proc. ICSLP*, vol. 4, pp. 1835–1838, 1994.

[60] T. Houtgast, H. J. M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function: I. General room acoustics," *Acustica*, no. 46, pp. 60–72, 1980.

[61] H. G. Hirsch and A. Corsten, "A new method to improve speech recognition in a noisy environment," *Signal Processing V – Proc. EUSIPCO*, pp. 1187–1190, Elsevier Science, 1990.

[62] H. Murveit, J. Butzburger, and M. Weintraub, "Reduced channel dependence for speech recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 280–284, Feb. 1992.

[63] J. Smolders and D. V. Compernolle, "In search for the relevant parameters for speaker independent speech recognition," *Proc. ICASSP*, vol. II, pp. 684–687, 1993.

[64] S. F. Boll, "Supression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol. 27, pp. 113–120, 1979.

[65] B. H. Juang and L. R. Rabiner, "Signal restoration by spectral mapping," *Proc. ICASSP*, pp. 2368–2371, 1987.

[66] M. J. F. Gales and S. J. Young, "Parallel model combination for speech recognition in additive and convolutional noise," CUED/F-INFENG/TR154, Cambridge U. Engineering Dept., Dec. 1993.

[67] D. Dubois, "Comparison of time-dependant acoustic features for a speaker-independent speech recognition system," *Proc. EUROSPEECH*, pp. 935–938, 1991.

[68] J.-C. Junqua, S. Valente, D. Fohr, and J.-F. Mari, "An N-best strategy, dynamic grammars and selectively trained neural networks for real-time recognition of continuously spelled names over the telephone," *Proc. ICASSP*, vol. 1, pp. 852–855, 1995.

[69] R. A. Cole, K. Roginski, and M. Fanty, "English alphabet recognition with telephone speech," *Proc. EUROSPEECH*, pp. 479–482, 1991.

[70] C. Nadeu and B.-H. Juang, "Filtering of spectral parameters for speech recognition," *Proc. ICSLP*, pp. 1927–1930, 1994.

[71] B. E. F. Lindblom and M. Studdert-Kennedy, "On the role of formant transitions in vowel recognition," *JASA*, vol. 42, pp. 830–843, 1967.

[72] M. J. Hunt and C. Lefèbvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," *Proc. ICASSP*, pp. 262–265, 1989.

[73] S. Furui, "Feature analysis based on articulatory and perceptual models," *Proc. IEEE Workshop on Automatic Speech Recognition*, pp. 63–64, 1993.